

TV-SVM: Total Variation Support Vector Machine for Semi-Supervised Data Classification

Xavier Bresson*

Ruiliang Zhang*

Abstract

We introduce semi-supervised data classification algorithms based on total variation (TV), Reproducing Kernel Hilbert Space (RKHS), support vector machine (SVM), Cheeger cut, labeled and unlabeled data points. We design binary and multi-class semi-supervised classification algorithms. We compare the TV-based classification algorithms with the related Laplacian-based algorithms, and show that TV classification perform significantly better when the number of labeled data is small.

1 Introduction

1.1 Notation

Let $\{x_i, y_i\}_{1 \leq i \leq N}$ denote N data points, where $x_i \in \mathbb{R}^d$ is its attributes with dimension d , while $y_i \in \{+1, -1\}$ (binary classification) or $y_i \in \{1, \dots, c\}$ (multi-class classification). The total number of data points is N including n labeled data and $N - n$ unlabeled data. H_K is a Reproducing Kernel Hilbert Space (RKHS) with $K : \mathbb{R}^{d \times d} \rightarrow \text{Sym}(\mathbb{R})$ is an operator-valued, positive definite kernel. Finally, we use the abbreviation $f_i = f(x_i)$.

2 Binary (two-class) data classification

2.1 Regularized Least Square (RLS)

The standard RLS problem for binary classification is as follows [8]. Find a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$\min_{f \in H_K} \frac{\eta}{2} \sum_{i \in n} (y_i - f_i)^2 + \frac{\lambda}{2} \|f\|_{H_K}^2, \quad (1)$$

where $\eta, \lambda > 0$. Representer theorem states the existence of a minimizing function $f^*(x) = \sum_{j \in n} K(x, x_j) \alpha_j^*$ (or $f(x) = K_x \alpha$ with matrix representation) and the norm of f in the RKHS is $\|f\|_{H_K}^2 = \alpha^T K \alpha$. Problem (1) is equivalent to

$$\min_{\alpha \in \mathbb{R}^n} \frac{\eta}{2} \|y - K\alpha\|_2^2 + \frac{\lambda}{2} \alpha^T K \alpha \quad (2)$$

Taking the derivative w.r.t. α provides the minimizer:

$$\alpha^* = (\eta K + \lambda I_n)^{-1} (\eta y) \quad (3)$$

Finally, unseen data points are classified as follows:

$$x \in C_1 \text{ if } f^*(x) \geq 0 \quad (4)$$

$$x \in C_2 \text{ if } f^*(x) < 0 \quad (5)$$

*Department of Computer Science, City University of Hong Kong, Hong Kong.

2.2 Laplacian-based RLS

The Laplacian-based RLS problem for binary semi-supervised classification is as follows [1]:

$$\min_{f \in H_K} \frac{\eta}{2} \sum_{i \in n} (y_i - f_i)^2 + \frac{\lambda}{2} \|f\|_{H_K}^2 + \underbrace{\frac{\gamma}{2} \sum_{i,j \in N} w_{i,j} |f_i - f_j|^2}_{\|Df\|^2}, \quad (6)$$

where $\|Df\|^2 = \sum_{i,j \in N} w_{i,j} |f_i - f_j|^2 = f^T L f$ is the Dirichlet energy and $L = D - W$ is the graph Laplacian. Observe that training data points are composed of n labeled points and $N - n$ unlabeled points. Let us consider matrix $J = \text{diag}(1, \dots, 1, 0, \dots, 0)$ with the first n diagonal entries as 1 and the rest 0 and $y = [y_1, \dots, y_n, 0, \dots, 0]$ with $N - n$ entries as 0. This allows to write $\sum_{i \in n} (y_i - f_i)^2 = \|y - Jf\|_2^2$. Representer theorem states the existence of a minimizing function $f^*(x) = \sum_{j \in N} K(x, x_j) \alpha_j^*$ exists. Problem (6) is equivalent to

$$\min_{\alpha \in \mathbb{R}^N} \frac{\eta}{2} \|y - JK\alpha\|_2^2 + \frac{\lambda}{2} \alpha^T K \alpha + \frac{\gamma}{2} (K\alpha)^T L (K\alpha). \quad (7)$$

Taking the derivative w.r.t. α provides the minimizer:

$$\alpha^* = (\eta JK + \lambda I_N + \gamma LK)^{-1} (\eta y) \quad (8)$$

Finally, unseen data points are classified as follows:

$$x \in C_1 \text{ if } f^*(x) \geq 0 \quad (9)$$

$$x \in C_2 \text{ if } f^*(x) < 0 \quad (10)$$

2.3 Total Variation-based RLS

The TV-based RLS problem for binary semi-supervised classification is as follows [6]:

$$\min_{f \in H_K} \frac{\eta}{2} \sum_{i \in n} (y_i - f_i)^2 + \frac{\lambda}{2} \|f\|_{H_K}^2 + \underbrace{\gamma \sum_{i,j \in N} w_{i,j} |f_i - f_j|}_{\|Df\|}, \quad (11)$$

where $\|Df\| = \sum_{i,j \in N} w_{i,j} |f_i - f_j|$ is the graph TV of function f . Unlike previous optimization problems, minimizing (11) needs advanced optimization techniques as TV term is non-differentiable. However, recent advances in ℓ^1 optimization provide efficient tools to deal with problem (11). In this work, we propose a splitting step coupled with an augmented Lagrangian method. Although one splitting variable is enough for minimizing (11), experimental observations suggest more accurate results using two splitting variables g, h . The proposed iterative optimization algorithm is as follows:

$$\begin{aligned} (f^{n+1}, h^{n+1}, g^{n+1}) &= \min_{f \in H_K, h, g} \frac{\eta}{2} \|y - Jh\|_2^2 + \frac{\lambda}{2} \|f\|_{H_K}^2 + \gamma \|Dg\| + \\ &< \lambda_1^n, f - g > + \frac{r_1}{2} \|f - g\|_2^2 + < \lambda_2^n, h - g > + \frac{r_2}{2} \|h - g\|_2^2 \end{aligned} \quad (12)$$

$$\lambda_1^{n+1} = \lambda_1^n + r_1 (f^{n+1} - g^{n+1}) \quad (13)$$

$$\lambda_2^{n+1} = \lambda_2^n + r_2 (h^{n+1} - g^{n+1}) \quad (14)$$

The sub-minimization problem w.r.t. f is:

$$\min_{f \in H_K} \frac{\lambda}{2} \|f\|_{H_K}^2 + \frac{r_1}{2} \|f - (g - \frac{\lambda_1}{r_1})\|_2^2 \quad (15)$$

which solution is given by $f^{n+1} = K\alpha^{n+1}$, with

$$\alpha^{n+1} = (\lambda I_N + r_1 K)^{-1} (r_1 g^n - \lambda_1^n) \quad (16)$$

The sub-minimization problem w.r.t. h is:

$$\min_h \frac{\eta}{2} \|y - Jh\|_2^2 + \frac{r_2}{2} \|h - (g - \frac{\lambda_2}{r_2})\|_2^2 \quad (17)$$

which solution is given by

$$h^{n+1} = (\eta J + r_2 I_N)^{-1} (\eta y + r_2 g^n - \lambda_2^n) \quad (18)$$

The sub-minimization problem w.r.t. g is:

$$\min_g \gamma \|Dg\| + \frac{r_1}{2} \|g - (f + \frac{\lambda_1}{r_1})\|_2^2 + \frac{r_2}{2} \|g - (h + \frac{\lambda_2}{r_2})\|_2^2 \quad (19)$$

which can be written as

$$\min_g \gamma \|Dg\| + \frac{r_1 + r_2}{2} \|g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2}\|_2^2 \quad (20)$$

with $z_1 = f + \frac{\lambda_1}{r_1}$ and $z_2 = h + \frac{\lambda_2}{r_2}$. Different techniques can be applied to solve the TV ROF problem [9]. We use the primal-dual method [3] which is guaranteed to converge in $O(\frac{1}{k^2})$, k being the iteration number. Finally, we project each function f, h, g on the unit ball (i.e. $f^{n+1} \leftarrow N \cdot \frac{f^{n+1}}{\|f^{n+1}\|_2}$) and constraint them to be zero-mean (i.e. $f^{n+1} \leftarrow f^{n+1} - \text{mean}(f^{n+1})$).

We summarize the iterative algorithm:

$$\alpha^{n+1} = (\lambda I_N + r_1 K)^{-1} (r_1 g^n - \lambda_1^n) \quad (21)$$

$$f^{n+1} = K \alpha^{n+1} \quad (22)$$

$$h^{n+1} = (\eta J + r_2 I_N)^{-1} (\eta y + r_2 g^n - \lambda_2^n) \quad (23)$$

$$\bar{g}^{n+1} = \text{argmin}_g \gamma \|Dg\| + \frac{r_1 + r_2}{2} \|g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2}\|_2^2 \quad (24)$$

$$\text{with } z_1 = f + \frac{\lambda_1^n}{r_1}, \quad z_2 = h + \frac{\lambda_2^n}{r_2} \quad (25)$$

$$\hat{g}^{n+1} = N \cdot \frac{\bar{g}^{n+1}}{\|\bar{g}^{n+1}\|_2} \quad (26)$$

$$g^{n+1} = \hat{g}^{n+1} - \text{mean}(\hat{g}^{n+1}) \quad (27)$$

2.4 Cheeger-based RLS

The Cheeger-based RLS problem for binary semi-supervised classification is as follows:

$$\min_{f \in H_K} \frac{\sum_{i,j \in N} w_{i,j} |f_i - f_j|}{\sum_{i \in N} |f_i - \text{median}(f)|} \quad \text{s.t. } f_i = y_i, \forall i \in n \quad (28)$$

Based on [2], the following algorithm is proposed:

$$g^{n+1} = f^n + c \cdot \text{sign}(f^n) \quad (29)$$

$$e^{n+1} = \text{RLS}(g^{n+1}) \quad (30)$$

$$h^{n+1} = \text{argmin}_h TV(h) + \frac{E^n}{2c} \|h - e^{n+1}\|_2^2 \quad (31)$$

$$t^{n+1} = h^{n+1} - \text{median}(h^{n+1}) \quad (32)$$

$$s^{n+1} = \begin{cases} y_i & \forall i \in n \\ t^{n+1}(i) & \forall i \notin n \end{cases} \quad (33)$$

$$f^{n+1} = N \cdot \frac{s^{n+1}}{\|s^{n+1}\|_2} \quad (34)$$

where $\text{RLS}(g)$ is as follow

$$\min_{e \in H_K} \frac{\lambda}{2} \|e\|_{H_K}^2 + \frac{r}{2} \|e - g\|_2^2, \quad (35)$$

which solution is given by $e^{n+1} = K\alpha^*$, with

$$\alpha^* = (\lambda I + rK)^{-1}rg. \quad (36)$$

2.5 Support Vector Machine (SVM)

The standard SVM method for binary classification is as follows [5]. Find a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ such that

$$\begin{aligned} \min_{f \in H_K, b \in \mathbb{R}} \quad & \frac{\lambda}{2} \|f\|_{H_K}^2, \\ \text{s.t. } & y_i(f_i + b) \geq 1, i = 1, \dots, n. \end{aligned} \quad (37)$$

where $\lambda > 0$. To deal with non-separable case, the above problem can be rewritten with a slack variable ξ :

$$\begin{aligned} \min_{f \in H_K, b \in \mathbb{R}, \xi \in \mathbb{R}^n} \quad & \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \sum_{i \in n} \xi_i, \\ \text{s.t. } & y_i(f_i + b) \geq 1 - \xi_i, i = 1, \dots, n, \\ & \xi_i \geq 0, i = 1, \dots, n \end{aligned} \quad (38)$$

Representer theorem states the existence of a minimizing function $f^*(x) = \sum_{j \in n} K(x, x_j)\alpha_j^*$ and $\|f\|_{H_K}^2 = \alpha^T K \alpha$. Problem (38) is equivalent to

$$\begin{aligned} \min_{\alpha, \xi \in \mathbb{R}^n, b \in \mathbb{R}} \quad & \frac{\lambda}{2} \alpha^T K \alpha + \mu \sum_{i=1}^n \xi_i, \\ \text{s.t. } & y_i \left(\sum_{j=1}^n K(x_i, x_j) \alpha_j + b \right) \geq 1 - \xi_i, i = 1, \dots, n, \\ & \xi_i \geq 0, i = 1, \dots, n \end{aligned} \quad (39)$$

By using the Lagrangian multiplier technique, problem (39) can be reformulated as:

$$\begin{aligned} \min_{\alpha, \xi, \beta, \beta_\xi \in \mathbb{R}^n, b \in \mathbb{R}} \quad & \frac{\lambda}{2} \alpha^T K \alpha + \mu \xi^T \mathbf{1} + \beta^T (\mathbf{1} - \xi - Y(K\alpha + b\mathbf{1})) - \beta_\xi^T \xi, \\ \text{s.t. } & \beta_i, \beta_{\xi_i} \geq 0, i = 1, \dots, n \end{aligned} \quad (40)$$

where β, β_ξ are Lagrangian multipliers, $\mathbf{1}$ is a vector whose elements are all ones, and $Y = \text{diag}(y_1, \dots, y_n)$. Let us consider the Lagrangian optimality conditions. Taking the derivative w.r.t. b and setting to 0 gives

$$\beta^T Y \mathbf{1} = 0 \Rightarrow \beta^T y = 0. \quad (41)$$

Taking the derivative w.r.t. ξ and setting to 0 gives

$$\mu \mathbf{1} - \beta - \beta_\xi = 0 \Rightarrow 0 \leq \beta_i \leq \mu, i = 1, \dots, n. \quad (42)$$

Taking the derivative w.r.t. α and setting to 0 gives

$$\alpha = \frac{Y\beta}{\lambda} \quad (43)$$

By substituting (43) back into (40), we reach the following dual optimization problem:

$$\begin{aligned} \max_{\beta \in \mathbb{R}^n} \quad & \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta, \\ \text{s.t. } & \beta^T y = 0, \\ & 0 \leq \beta_i \leq \mu, i = 1, \dots, n \end{aligned} \quad (44)$$

where $Q = Y(\frac{K}{\lambda})Y$. The above problem can be solved using several efficient SVM solvers s.a. libSVM [4]. Once the optimal β^* is obtained, it is straightforward to get the optimal α^* :

$$\alpha^* = \frac{Y\beta^*}{\lambda} \quad (45)$$

and

$$f^*(x) = \sum_{i=1}^n \alpha_i^* K(x, x_i). \quad (46)$$

Finally, unseen data points are classified as follows:

$$x \in C_1 \text{ if } f^*(x) \geq 0 \quad (47)$$

$$x \in C_2 \text{ if } f^*(x) < 0 \quad (48)$$

2.6 Laplacian-based SVM

The Laplacian-based SVM problem with slack variable for binary semi-supervised classification is as follows [1]:

$$\begin{aligned} \min_{f \in H_K, \xi \in \mathbb{R}^N, b \in \mathbb{R}} \quad & \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \sum_{i \in N} \xi_i + \frac{\gamma}{2} \underbrace{\sum_{i,j \in N} w_{i,j} |f_i - f_j|^2}_{\|Df\|^2}, \\ \text{s.t. } & y_i(f_i + b) \geq 1 - \xi_i, i = 1, \dots, N \\ & \xi_i \geq 0, i = 1, \dots, N \end{aligned} \quad (49)$$

By using Lagrangian multipliers technique, problem (49) becomes:

$$\begin{aligned} \min_{\alpha, \xi, \beta, \beta_{\xi} \in \mathbb{R}^N, b \in \mathbb{R}} \quad & \frac{\lambda}{2} \alpha^T K \alpha + \mu \xi^T \mathbf{1} + \frac{\gamma}{2} \alpha^T K L K \alpha + \beta^T (\mathbf{1} - \xi - Y(K\alpha + b\mathbf{1})) - \beta_{\xi}^T \xi, \\ \text{s.t. } & \beta_i, \beta_{\xi_i} \geq 0, i = 1, \dots, N \end{aligned} \quad (50)$$

Applying the same steps as (41),(42) and (43), we get

$$\begin{aligned} \max_{\beta \in \mathbb{R}^N} \quad & \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta, \\ \text{s.t. } & \beta^T y = 0, \\ & 0 \leq \beta_i \leq \mu, i = 1, \dots, N \end{aligned} \quad (51)$$

where

$$Q = Y(\lambda I + \gamma L K)^{-1} K Y \quad (52)$$

Optimal α^* is obtained by solving the following linear system:

$$\alpha^* = (\lambda I + \gamma L K)^{-1} Y \beta^* \quad (53)$$

and

$$f^*(x) = \sum_{i=1}^N \alpha_i^* K(x, x_i). \quad (54)$$

Finally, unseen data points are classified as follows:

$$x \in C_1 \text{ if } f^*(x) \geq 0 \quad (55)$$

$$x \in C_2 \text{ if } f^*(x) < 0 \quad (56)$$

2.7 Total Variation-based SVM

The TV-based SVM for binary semi-supervised classification is as follows:

$$\begin{aligned}
\min_{f \in H_K, \xi \in \mathbb{R}^N, b \in \mathbb{R}} \quad & \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \sum_{i \in N} \xi_i + \gamma \underbrace{\sum_{i,j \in N} w_{i,j} |f_i - f_j|}_{\|Df\|}, \\
\text{s.t.} \quad & y_i(f_i + b) \geq 1 - \xi_i, i = 1, \dots, N \\
& \xi_i \geq 0, i = 1, \dots, N
\end{aligned} \tag{57}$$

where $\|Df\| = \sum_{i,j \in N} w_{i,j} |f_i - f_j|$ is the graph TV of function f . Like for TV-based RLS, we use a splitting step coupled with an augmented Lagrangian method. The proposed iterative optimization algorithm is as follows:

$$\begin{aligned}
(f^{n+1}, h^{n+1}, g^{n+1}) = \min_{f \in H_K, h, g, \xi, b} \quad & \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \sum_{i=1}^N \xi_i + \gamma \|Dg\| + \\
& \langle \lambda_1^n, f - g \rangle + \frac{r_1}{2} \|f - g\|_2^2 + \langle \lambda_2^n, h - g \rangle + \frac{r_2}{2} \|h - g\|_2^2
\end{aligned} \tag{58}$$

$$\text{s.t.} \quad y_i(h_i + b) \geq 1 - \xi_i, i = 1, \dots, N \tag{59}$$

$$\begin{aligned}
& \xi_i \geq 0, i = 1, \dots, N \\
& \lambda_1^{n+1} = \lambda_1^n + r_1(f^{n+1} - g^{n+1}) \\
& \lambda_2^{n+1} = \lambda_2^n + r_2(h^{n+1} - g^{n+1})
\end{aligned} \tag{60}$$

$$\lambda_1^{n+1} = \lambda_1^n + r_1(f^{n+1} - g^{n+1}) \tag{61}$$

$$\lambda_2^{n+1} = \lambda_2^n + r_2(h^{n+1} - g^{n+1}) \tag{62}$$

The sub-minimization problem w.r.t. f is:

$$\min_{f \in H_K} \frac{\lambda}{2} \|f\|_{H_K}^2 + \frac{r_1}{2} \|f - (g - \frac{\lambda_1}{r_1})\|_2^2 \tag{63}$$

which solution is given by $f^{n+1} = K\alpha^{n+1}$, with

$$\alpha^{n+1} = (\lambda I_N + r_1 K)^{-1} (r_1 g^n - \lambda_1^n) \tag{64}$$

The sub-minimization problem w.r.t. h is:

$$\begin{aligned}
\min_{h, \xi, b} \quad & \mu \sum_{i=1}^N \xi_i + \frac{r_2}{2} \|h - e\|_2^2 \\
\text{s.t.} \quad & y_i(h_i + b) \geq 1 - \xi_i, i = 1, \dots, N \\
& \xi_i \geq 0, i = 1, \dots, N
\end{aligned} \tag{65}$$

where $e = g - \frac{\lambda_2}{r_2}$. Problem (65) is equivalent to

$$\begin{aligned}
\min_{h, \xi, b, \beta, \beta_\xi} \quad & \mu \xi^T \mathbf{1} + \frac{r_2}{2} \|h - e\|_2^2 + \beta^T (\mathbf{1} - \xi - Y(h + b)) - \beta_\xi^T \xi \\
\text{s.t.} \quad & \beta_i, \beta_{\xi i} \geq 0, i = 1, \dots, N
\end{aligned} \tag{66}$$

Applying the same steps as (41),(42) and (43) did, we get

$$\begin{aligned}
\max_{\beta \in \mathbb{R}^N} \quad & \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \beta^T P, \\
\text{s.t.} \quad & \beta^T y = 0, \\
& 0 \leq \beta_i \leq \mu, i = 1, \dots, N
\end{aligned} \tag{67}$$

where $Q = \frac{YY^T}{r_2}$ and $P = Ye$. Problem (67) can be solved by gradient descent method, and the solution β^* can be used to obtain the optimal h^{n+1} :

$$h^{n+1} = \frac{1}{r_2} Y \beta^* + e \tag{68}$$

The sub-minimization problem w.r.t. g is:

$$\min_g \gamma \|Dg\| + \frac{r_1}{2} \|g - (f + \frac{\lambda_1}{r_1})\|_2^2 + \frac{r_2}{2} \|g - (h + \frac{\lambda_2}{r_2})\|_2^2 \quad (69)$$

which can be written as

$$\min_g \gamma \|Dg\| + \frac{r_1 + r_2}{2} \|g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2}\|_2^2 \quad (70)$$

with $z_1 = f + \frac{\lambda_1}{r_1}$ and $z_2 = h + \frac{\lambda_2}{r_2}$.

We summarize the iterative algorithm:

$$\alpha^{n+1} = (\lambda I_N + r_1 K)^{-1} (r_1 g^n - \lambda_1^n) \quad (71)$$

$$f^{n+1} = K \alpha^{n+1} \quad (72)$$

$$\beta^* = \max_{\beta \in \mathbb{R}^N} \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \beta^T P, \text{ s.t. } \beta^T y = 0, 0 \leq \beta_i \leq C, i = 1, \dots, N \quad (73)$$

$$h^{n+1} = \frac{1}{r_2} Y \beta^* + e \quad (74)$$

$$\bar{g}^{n+1} = \operatorname{argmin}_g \gamma \|Dg\| + \frac{r_1 + r_2}{2} \|g - \frac{r_1 z_1 + r_2 z_2}{r_1 + r_2}\|_2^2 \quad (75)$$

$$\text{with } z_1 = f + \frac{\lambda_1^n}{r_1}, z_2 = h + \frac{\lambda_2^n}{r_2} \quad (76)$$

$$\hat{g}^{n+1} = N \cdot \frac{\bar{g}^{n+1}}{\|\bar{g}^{n+1}\|_2} \quad (77)$$

$$g^{n+1} = \hat{g}^{n+1} - \operatorname{mean}(\hat{g}^{n+1}) \quad (78)$$

2.8 Cheeger-based SVM

The Cheeger-based SVM problem for binary semi-supervised classification is as follows:

$$\min_{f \in H_K} \frac{\sum_{i,j \in N} w_{i,j} |f_i - f_j|}{\sum_{i \in N} |f_i - \operatorname{median}(f)|} \text{ s.t. } f_i = y_i, \forall i \in n \quad (79)$$

$$\text{s.t. } y_i(f_i + b) \geq 1 - \xi_i, i = 1, \dots, N \quad (80)$$

$$\xi_i \geq 0, i = 1, \dots, N \quad (81)$$

Based on [2], the following algorithm is proposed:

$$g^{n+1} = f^n + c \cdot \operatorname{sign}(f^n) \quad (82)$$

$$e^{n+1} = \operatorname{SVM}(g^{n+1}) \quad (83)$$

$$h^{n+1} = \operatorname{argmin}_h TV(h) + \frac{E^n}{2c} \|h - e^{n+1}\|_2^2 \quad (84)$$

$$t^{n+1} = h^{n+1} - \operatorname{median}(h^{n+1}) \quad (85)$$

$$s^{n+1} = \begin{cases} l(i) & \forall i \in n \\ t^{n+1}(i) & \forall i \notin n \end{cases} \quad (86)$$

$$f^{n+1} = N \cdot \frac{s^{n+1}}{\|s^{n+1}\|_2} \quad (87)$$

where $\operatorname{SVM}(g)$ is as follow:

$$\begin{aligned} \min_{e, \xi, b} \quad & \frac{\lambda}{2} \|e\|_{H_K}^2 + \mu \sum_{i \in N} \xi_i + \frac{r}{2} \|e - g\|_2^2 \\ \text{s.t.} \quad & y_i(e_i + b) \geq 1 - \xi_i, i = 1, \dots, N \\ & \xi_i \geq 0, i = 1, \dots, N \end{aligned} \quad (88)$$

Problem (88) is equivalent to

$$\begin{aligned} \min_{e, \xi, b, \beta, \beta_\xi} \quad & \frac{\lambda}{2} \|e\|_{H_K}^2 + \mu \xi^T \mathbf{1} + \frac{r}{2} \|e - g\|_2^2 + \beta^T (\mathbf{1} - \xi - Y(e + b)) - \beta_\xi^T \xi \\ \text{s.t.} \quad & \beta_i, \beta_{\xi_i} \geq 0, i = 1, \dots, N \end{aligned} \quad (89)$$

Applying the same steps as (41), (42) and (43), we get

$$\begin{aligned} \max_{\beta \in \mathbb{R}^N} \quad & \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \frac{1}{2} P \beta, \\ \text{s.t.} \quad & \beta^T y = 0, \\ & 0 \leq \beta_i \leq \mu, i = 1, \dots, N \end{aligned} \quad (90)$$

where $Q = YGY$, $P = rg^T(G + G^T)Y$ and $G = (\lambda I + rK)^{-1}K$. The above problem can be solved by gradient descent method, and the solution β^* can be used to obtain the optimal α^* :

$$\alpha^* = (\lambda I + rK)^{-1}(Y\beta^* + rg) \quad (91)$$

and

$$e^{n+1} = K\alpha^* \quad (92)$$

2.9 Experimental results

# labels per class	1	5	10	50
Lap-RLS	18.09	10.48	7.77	4.14
Lap-SVM	13.79	9.84	7.61	4.77
TV-RLS	3.18	3.16	3.13	3.16
TV-SVM	3.18	3.13	3.13	3.08
Cheeger-RLS	4.06	3.74	4.03	2.84
Cheeger-SVM	3.87	3.74	4.00	2.73

Table 1: Binary semi-supervised classification algorithms tested on the sets of 4's and 9's from USPS dataset. Error is averaged over 10 runs with randomly selected labels.

3 Multi-class data classification

3.1 Laplacian-based RLS

The Laplacian-based RLS problem for multi-class semi-supervised classification is as follows:

$$\begin{aligned} \min_{\vec{f}=(f^1, \dots, f^c) \in H_K} \quad & \frac{\eta}{2} \sum_{k=1}^c \sum_{i \in n} (y_i^k - f_i^k)^2 + \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \underbrace{\frac{\gamma}{2} \sum_{k=1}^c \sum_{i,j \in N} w_{i,j} |f_i^k - f_j^k|^2}_{\|Df^k\|^2}, \\ \text{s.t.} \quad & \sum_{k=1}^c f_i^k = 1, f_i^k \geq 0, \forall i \in N \end{aligned} \quad (93)$$

where the last constraint being the simplex constraint. Problem (93) is equivalent to

$$\begin{aligned} \min_{\vec{f}=(f^1, \dots, f^c) \in H_K} \quad & \frac{\eta}{2} \sum_{k=1}^c \sum_{i \in n} (y_i^k - f_i^k)^2 + \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \frac{\gamma}{2} \sum_{k=1}^c \|Df^k\|^2, \\ \text{s.t.} \quad & f^k = g^k, \sum_{k=1}^c g_i^k = 1, g_i^k \geq 0, \forall i \in N \end{aligned} \quad (94)$$

This leads to the proposed iterative algorithm:

$$(\alpha^k)^{n+1} = \underset{\alpha^k \in \mathbb{R}^N}{\operatorname{argmin}} \frac{\eta}{2} \|y^k - J^k K \alpha^k\|_2^2 + \frac{\lambda}{2} \alpha_k^T K \alpha_k + \frac{\gamma}{2} (K \alpha_k)^T L (K \alpha_k) + \frac{r}{2} \|K \alpha^k - (g^k - \frac{\lambda^k}{r})\|_2^2 \quad (95)$$

$$= (\eta J^k K + r K + \lambda I_N + \gamma L K)^{-1} (\eta y^k + r g^k - \lambda^k) \quad (96)$$

$$(f^k)^{n+1} = K(\alpha^k)^{n+1} \quad (97)$$

$$(g^k)^{n+1} = \Pi_{\sum g^k=1} (f^k + \frac{\lambda^k}{r}) \quad (98)$$

The simplex projection is done by Michelot's method [7].

Finally, unseen data points are classified as follows:

$$x \in C_k \text{ if } f_k^*(x) = \max_j (\{f_j^*(x)\}_{1 \leq j \leq c}) \quad (99)$$

3.2 Total Variation-based RLS

The TV-based RLS problem for multi-class semi-supervised classification is as follows:

$$\begin{aligned} \min_{\bar{f}=(f^1, \dots, f^c) \in H_K} \quad & \frac{\eta}{2} \sum_{k=1}^c \sum_{i \in n} (y_i^k - f_i^k)^2 + \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \gamma \underbrace{\sum_{i,j \in N} w_{i,j} |f_i^k - f_j^k|}_{\|Df^k\|} \\ \text{s.t.} \quad & \sum_{k=1}^c f^k(i) = 1, \quad f^k(i) \geq 0, \forall i \in N \end{aligned} \quad (100)$$

Problem (100) is equivalent to

$$\begin{aligned} \min_{\bar{f}=(f^1, \dots, f^c) \in H_K} \quad & \frac{\eta}{2} \sum_{k=1}^c \sum_{i \in L} (y_i^k - f_i^k)^2 + \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \frac{\gamma}{2} \sum_{k=1}^c \|Df^k\|, \\ \text{s.t.} \quad & f^k = g^k, \quad \sum_{k=1}^c g_i^k = 1, \quad g_i^k \geq 0, \forall i \in N \end{aligned} \quad (101)$$

This leads to the proposed iterative algorithm:

$$(\alpha^k)^{n+1} = \underset{\alpha^k \in \mathbb{R}^N}{\operatorname{argmin}} \frac{\eta}{2} \|y^k - J^k K \alpha^k\|_2^2 + \frac{\lambda}{2} \alpha_k^T K \alpha_k + \frac{r}{2} \|K \alpha^k - (g^k - \frac{\lambda^k}{r})\|_2^2 \quad (102)$$

$$= (\eta J^k K + r K + \lambda I_N)^{-1} (\eta y^k + r g^k - \lambda^k) \quad (103)$$

$$(f^k)^{n+1} = K(\alpha^k)^{n+1} \quad (104)$$

$$(\hat{g}^k)^{n+1} = \underset{g^k}{\operatorname{argmin}} \gamma \|Dg^k\| + \frac{r}{2} \|g^k - (f^k + \frac{\lambda^k}{r})\|_2^2 \quad (105)$$

$$(\bar{g}^k)^{n+1} = \Pi_{\sum g^k=1} (\hat{g}^k) \quad (106)$$

$$(g^k)^{n+1} = N \cdot \frac{(\bar{g}^k)^{n+1}}{\|(\bar{g}^k)^{n+1}\|_2} \quad (107)$$

Finally, unseen data points are classified as follows:

$$x \in C_k \text{ if } f_k^*(x) = \max_j (\{f_j^*(x)\}_{1 \leq j \leq c}) \quad (107)$$

3.3 Cheeger-based RLS

The Cheeger-based RLS problem for multi-class semi-supervised classification is as follows:

$$\min_{\vec{f}=(f^1,\dots,f^c)\in H_K} \sum_{k=1}^c \frac{\sum_{i,j\in N} w_{i,j} |f_i^k - f_j^k|}{\sum_{i\in N} |f_i^k - \text{median}(f^k)|} \text{ s.t. } f_i^k = l_i^k, \forall i \in n \quad (108)$$

$$(109)$$

The following algorithm is proposed:

$$(g^k)^{n+1} = (f^k)^n + c \cdot \text{sign}((f^k)^n) \quad (110)$$

$$(e^k)^{n+1} = \text{RLS}((g^k)^{n+1}) \quad (111)$$

$$(h^k)^{n+1} = \text{argmin}_{h^k} TV(h^k) + \frac{E^n}{2c} \|h^k - (e^k)^{n+1}\|_2^2 \quad (112)$$

$$(t^k)^{n+1} = (h^k)^{n+1} - \text{median}((h^k)^{n+1}) \quad (113)$$

$$(s^k)^{n+1} = \begin{cases} y_i^k & \forall i \in n \\ (t^k)^{n+1}(i) & \forall i \notin n \end{cases} \quad (114)$$

$$(\hat{s}^k)^{n+1} = \Pi_{\sum s^k=1}(s^k) \quad (115)$$

$$(f^k)^{n+1} = N \cdot \frac{(\hat{s}^k)^{n+1}}{\|(\hat{s}^k)^{n+1}\|_2} \quad (116)$$

where $\text{RLS}(g)$ is exact the same as (35).

Finally, unseen data points are classified as follows:

$$x \in C_k \text{ if } f_k^*(x) = \max_j (\{f_j^*(x)\}_{1 \leq j \leq c}) \quad (117)$$

3.4 Laplacian-based SVM

The Laplacian-based SVM for multi-class semi-supervised classification is as follows:

$$\begin{aligned} \min_{\vec{f}=(f^1,\dots,f^c)\in H_K, b\in\mathbb{R}^c, \xi\in\mathbb{R}^{N\times c}} & \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \mu \sum_{k=1}^c \sum_{i\in N} \xi_i^k + \frac{\gamma}{2} \underbrace{\sum_{k=1}^c \sum_{i,j\in N} w_{i,j} |f_i^k - f_j^k|^2}_{\|Df^k\|^2}, \\ \text{s.t. } & y_i^k (f_i^k + b^k) \geq 1 - \xi_i^k, \xi_i^k \geq 0, i \in N, k \in c \\ & \sum_{k=1}^c f_i^k = 1, f^k(i) \geq 0, \forall i \in N \end{aligned}$$

Problem (118) is equivalent to

$$\begin{aligned} \min_{\vec{f}=(f^1,\dots,f^c)\in H_K, b\in\mathbb{R}^c, \xi\in\mathbb{R}^{N\times c}} & \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \mu \sum_{k=1}^c \sum_{i\in N} \xi_i^k + \frac{\gamma}{2} \sum_{k=1}^c \|Df^k\|^2, \\ \text{s.t. } & y_i^k (f_i^k + b^k) \geq 1 - \xi_i^k, \xi_i^k \geq 0, i \in N, k \in c \\ & f^k = g^k, \sum_{k=1}^c g_i^k = 1, g_i^k \geq 0, \forall i \in N \end{aligned}$$

Notes that, each f^k can be solved independently by using the same procedure as below (superscript k is ignored for convenience):

$$\begin{aligned} \min_{f\in H_K, b\in\mathbb{R}, \xi\in\mathbb{R}^N} & \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \xi^T \mathbf{1} + \frac{\gamma}{2} f^T L f + \frac{r}{2} \|f - e\|_2^2, \\ \text{s.t. } & y_i (f_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i \in N \end{aligned} \quad (118)$$

where $e = g - \frac{l}{r}$, and l is the Lagrangian multiplier. Problem (118) is equivalent to

$$\begin{aligned} \min_{b \in \mathbb{R}, \alpha, \xi, \beta, \beta_\xi \in \mathbb{R}^N} & \frac{\lambda}{2} \alpha^T K \alpha + \mu \xi^T \mathbf{1} + \frac{\gamma}{2} \alpha^T K L K \alpha + \frac{r}{2} \|K \alpha - e\|_2^2 \\ & + \beta^T (\mathbf{1} - \xi - Y(K \alpha + b)) - \beta_\xi^T \xi \\ \text{s.t. } & \beta, \beta_\xi \geq 0, i \in N \end{aligned} \quad (119)$$

Applying the same steps as (41), (42) and (43), we get

$$\begin{aligned} \max_{\beta \in \mathbb{R}^N} & \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \frac{1}{2} P \beta, \\ \text{s.t. } & \beta^T y = 0, \\ & 0 \leq \beta_i \leq \mu, i = 1, \dots, N \end{aligned} \quad (120)$$

where $Q = YGY$, $P = re^T(G + G^T)Y$ and $G = (\lambda I + \gamma LK + rK)^{-1}K$. The above problem can be solved by gradient descent method, and the solution β^* can be used to obtain the optimal α^* , which is:

$$\alpha^* = (\lambda I + \gamma LK + rK)^{-1}(Y\beta^* + re) \quad (121)$$

and

$$f = K\alpha^* \quad (122)$$

This leads to the following iterative algorithm:

$$(f^k)^{n+1} = \text{computed by using (120), (121) and (122)} \quad (123)$$

$$(g^k)^{n+1} = \Pi_{\sum g^k=1}((f^k)^{n+1} + \frac{l^k}{r}). \quad (124)$$

The simplex projection is done by Michelot's method [7].

Finally, unseen data points are classified as follows:

$$x \in C_k \text{ if } f_k^*(x) = \max_j \{f_j^*(x)\}_{1 \leq j \leq c} \quad (125)$$

3.5 Total Variation-based SVM

The TV-based SVM for multi-class semi-supervised classification is as follows:

$$\begin{aligned} \min_{\vec{f}=(f^1, \dots, f^c) \in H_K, \xi \in \mathbb{R}^{N \times c}, b \in \mathbb{R}^c} & \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \mu \sum_{k=1}^c \sum_{i=1}^N \xi_i^k + \gamma \underbrace{\sum_{k=1}^c \sum_{i,j \in N} w_{i,j} |f_i^k - f_j^k|}_{\|Df^k\|} \\ \text{s.t. } & y_i^k (f_i^k + b^k) \geq 1 - \xi_i^k, \xi_i^k \geq 0, i \in N, k \in c \end{aligned} \quad (126)$$

Problem (126) is equivalent to

$$\begin{aligned} \min_{\vec{f}=(f^1, \dots, f^c) \in H_K, \xi \in \mathbb{R}^{N \times c}, b \in \mathbb{R}^c} & \frac{\lambda}{2} \sum_{k=1}^c \|f^k\|_{H_K}^2 + \mu \sum_{k=1}^c \sum_{i=1}^N \xi_i^k + \gamma \sum_{k=1}^c \|Dg^k\|, \\ \text{s.t. } & y_i^k (f_i^k + b^k) \geq 1 - \xi_i^k, \xi_i^k \geq 0, i \in N, k \in c \\ & f^k = g^k, \sum_{k=1}^c g^k(i) = 1, g^k(i) \geq 0, \forall i \in N \end{aligned} \quad (127)$$

Notes that, each f^k can be solved independently:

$$\begin{aligned} \min_{f \in H_K, b \in \mathbb{R}, \xi \in \mathbb{R}^N} & \frac{\lambda}{2} \|f\|_{H_K}^2 + \mu \xi^T \mathbf{1} + \frac{r}{2} \|f - e\|_2^2, \\ \text{s.t. } & y_i (f_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i \in N \end{aligned} \quad (128)$$

where $e = g - \frac{l}{r}$, and l is the Lagrangian multiplier. Problem (128) is equivalent to

$$\begin{aligned} \min_{b \in \mathbb{R}, \alpha, \xi, \beta, \beta_\xi \in \mathbb{R}^N} \quad & \frac{\lambda}{2} \alpha^T K \alpha + \mu \xi^T \mathbf{1} + \frac{r}{2} \|K \alpha - e\|_2^2 + \beta^T (\mathbf{1} - \xi - Y(K \alpha + b)) - \beta_\xi^T \xi \\ \text{s.t.} \quad & \beta, \beta_\xi \geq 0, i \in N \end{aligned} \quad (129)$$

Applying the same steps as (41), (42) and (43), we get

$$\begin{aligned} \max_{\beta \in \mathbb{R}^N} \quad & \beta^T \mathbf{1} - \frac{1}{2} \beta^T Q \beta - \frac{1}{2} P \beta, \\ \text{s.t.} \quad & \beta^T y = 0, \\ & 0 \leq \beta_i \leq \mu, i = 1, \dots, N \end{aligned} \quad (130)$$

where $Q = YGY$, $P = re^T(G + G^T)Y$ and $G = (\lambda I + rK)^{-1}K$. The above problem can be solved by gradient descent method, and the solution β^* can be used to obtain the optimal α^* , which is:

$$\alpha^* = (\lambda I + rK)^{-1}(Y\beta^* + re) \quad (131)$$

and

$$f = K\alpha^* \quad (132)$$

This leads to the proposed iterative algorithm:

$$(f^k)^{n+1} = \text{computed by using (130), (131) and (132)} \quad (133)$$

$$(\hat{g}^k)^{n+1} = \operatorname{argmin}_{g^k} \gamma \|Dg^k\| + \frac{r}{2} \|g^k - (f^k + \frac{l^k}{r})\|_2^2 \quad (134)$$

$$(\bar{g}^k)^{n+1} = \Pi_{\sum g^k=1}(\hat{g}^k) \quad (135)$$

$$(g^k)^{n+1} = N \cdot \frac{(\bar{g}^k)^{n+1}}{\|(\bar{g}^k)^{n+1}\|_2} \quad (136)$$

Finally, unseen data points are classified as follows:

$$x \in C_k \text{ if } f_k^*(x) = \max_j \{f_j^*(x)\}_{1 \leq j \leq c} \quad (137)$$

3.6 Cheeger-based SVM

The Cheeger-based SVM with slack variable problem for multi-class classification is as follows:

$$\min_{\vec{f}=(f^1, \dots, f^c) \in H_K} \sum_{k=1}^c \frac{\sum_{i,j \in N} w_{i,j} |f_i^k - f_j^k|}{\sum_{i \in N} |f_i^k - \operatorname{median}(f^k)|} \text{ s.t. } f_i^k = y_i^k, \forall i \in n \quad (138)$$

The following algorithm is proposed:

$$(g^k)^{n+1} = (f^k)^n + c \cdot \operatorname{sign}((f^k)^n) \quad (139)$$

$$(e^k)^{n+1} = \operatorname{SVM}((g^k)^{n+1}) \quad (140)$$

$$(h^k)^{n+1} = \operatorname{argmin}_{h^k} TV(h^k) + \frac{E^n}{2c} \|h^k - (e^k)^{n+1}\|_2^2 \quad (141)$$

$$(t^k)^{n+1} = (h^k)^{n+1} - \operatorname{median}((h^k)^{n+1}) \quad (142)$$

$$(s^k)^{n+1} = \begin{cases} y^k(i) & \forall i \in n \\ (t^k)^{n+1}(i) & \forall i \notin n \end{cases} \quad (143)$$

$$(\hat{s}^k)^{n+1} = \Pi_{\sum s^k=1}(s^k) \quad (144)$$

$$(f^k)^{n+1} = N \cdot \frac{(\hat{s}^k)^{n+1}}{\|(\hat{s}^k)^{n+1}\|_2} \quad (145)$$

where $\operatorname{SVM}(\cdot)$ is as (88).

Finally, unseen data points are classified as follows:

$$x \in C_k \text{ if } f_k^*(x) = \max_j \{f_j^*(x)\}_{1 \leq j \leq c} \quad (146)$$

3.7 Experimental results

# labels per class	1	5	10	50
Lap-RLS	20.06	6.64	4.03	3.3
Lap-SVM	49.95	14.21	6.27	2.82
TV-RLS	2.0	2.06	1.91	1.98
TV-SVM	1.75	1.82	1.77	1.85
Cheeger-RLS	3.35	1.95	1.85	1.87
Cheeger-SVM	2.94	2.08	1.72	1.74

Table 2: Multi-class semi-supervised classification algorithms tested on four classes (0’s, 1’s, 4’s and 9’s) from USPS dataset. Error is averaged over 10 runs with randomly selected labels.

References

- [1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *The Journal of Machine Learning Research*, 7:2399–2434, 2006.
- [2] X. Bresson, T. Laurent, D. Uminsky, and J.H. von Brecht. Convergence and Energy Landscape for Cheeger Cut Clustering. *Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.
- [3] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.
- [4] C.C. Chang and C.J. Lin. LIBSVM: a Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 27(3), 2011.
- [5] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
- [6] S.H. Kang, B. Shafei, and G. Steidl. Supervised and Transductive Multi-Class Segmentation Using p-Laplacians and RKHS methods. 2012.
- [7] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of n . *Journal of Optimization Theory and Applications*, 50(1):195–200, 1986.
- [8] T. Poggio and S. Smale. The Mathematics of Learning: Dealing with Data. *Notices of the AMS*, 50(5):537–544, 2003.
- [9] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear Total Variation Based Noise Removal Algorithms. *Physica D*, 60(1-4):259 – 268, 1992.